University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2016

Leveraging Biometric Data to Boost Software Developer Productivity

Fritz, Thomas ; Müller, Sebastian

Abstract: Producing great software requires great productive developers. Yet, what does it really mean for an individual developer to be productive, and what can we do to best help developers to be productive? To answer these questions, research has traditionally focused on measuring a developer's output and therefore suffered from two drawbacks: the measures can only be calculated after a developer finished her work and these measures do not account for individual differences between developers. The recent advances in biometric sensor technology offer new opportunities to measure a developer's cognitive and emotional states in real-time and thus allow us to know more about what an individual developer is currently experiencing and what might foster or impede the developer's productivity. Results from recent research studies demonstrate the potential that biometric data has to accurately predict aspects of a developer's work, such as perceived task and code difficulty, progress and interruptibility of a developer. This opens up new opportunities for better supporting developers in their work and, for instance, prevent bugs from entering the code, reduce costly interruptions, and foster a better and more productive work day. Our vision is that biometric sensing will be integrated into a developer's work and that biometrics can be

DOI: https://doi.org/10.1109/SANER.2016.107

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-123101
Conference or Workshop Item

Originally published at:
Fritz, Thomas; Müller, Sebastian (2016). Leveraging Biometric Data to Boost Software Developer Productivity. In: International Conference on Software Analysis, Evolution and Reengeneering (Future of Software Engineering Track), Osaka, Japan, 14 March 2016 - 18 March 2016, s.n..
DOI: https://doi.org/10.1109/SANER.2016.107

Leveraging Biometric Data to Boost Software Developer Productivity

Thomas Fritz and Sebastian C. Müller
Department of Informatics
University of Zurich, Switzerland
Email: {fritz,smueller}@ifi.uzh.ch

***Abstract*—Producing great software requires great productive developers. Yet, what does it really mean for an individual developer to be productive, and what can we do to best help developers to be productive? To answer these questions, research has traditionally focused on measuring a developer's output and therefore suffered from two drawbacks: the measures can only be calculated after a developer finished her work and these measures do not account for individual differences between developers.**

The recent advances in biometric sensor technology offer new opportunities to measure a developer's cognitive and emotional states in real-time and thus allow us to know more about what an individual developer is currently experiencing and what might foster or impede the developer's productivity. Results from recent research studies demonstrate the potential that biometric data has to accurately predict aspects of a developer's work, such as perceived task and code difficulty, progress and interruptibility of a developer. This opens up new opportunities for better supporting developers in their work and, for instance, prevent bugs from entering the code, reduce costly interruptions, and foster a better and more productive work day. Our vision is that biometric sensing will be integrated into a developer's work and that biometrics can be used to boost the productivity of each individual developer.

I. Introduction

"Software is eating the world." This statement by Marc Andreesen in a Wall Street Journal Op-Ed highlights the fact that software has become the backbone of countless major businesses, a trend that is likely to continue for the foreseeable future [1]. Yet, there never seem to be enough software developers to satisfy the demand, despite the immense growth in the number of software developers over the years, with an estimate of eleven million professional developers in 2014 [2]. Beyond simply training more developers, one promising and complementary way to address this demand is to unleash the untapped potential of each "individual" developer. This raises some intriguing unanswered questions: What does it really mean for an individual developer to be productive? How are developers doing their work, what is going on in their minds and when are they experiencing difficulties? What are the biggest impediments to a developer's productivity and how can we best help to increase a developer's productivity?

Traditionally, researchers aiming to increase developers' productivity, have focused on what developers have done, measuring their output and collecting data from software repositories. For instance, several approaches have been developed to automatically detect defects in the code based on metrics such as code churn or code complexity [3], [4]. These approaches can help in decreasing maintenance cost and effort, however, they have two general downsides: first, the used metrics can only be calculated after change tasks are completed and second, they do not take into account the individual differences between developers, such as the experience level. With the Personal Software Process (PSP) Humphrey, as one of few, focused more on the individual by helping them to improve their skills and quality of work, but PSP requires developers to track measures manually, such as their schedule data, which is cumbersome and only allows more coarse granular insights [5].

Emerging biometric (aka. psychophysiological) sensor technology offers new ways to measure more of a developer, such as her cognitive and emotional states while she is working, rather than just the outputs of her work. The idea behind most biometric sensor technology is to measure physiological features of a person that can in turn be linked to the person's psychological states. As an example, a person who is stressed generally tends to sweat more than in less stressful situations and this difference in sweat leads to a varying electronic conductance of the skin that can be measured by electro-dermal activity (EDA) sensors. Extensive research in psychology has already investigated and correlated biometric measures, including skin-, heart-, eye- and brain-related ones, with a person's cognitive and emotional states. For instance, researchers have found that brain- and skin-related measurements can be linked to mental and cognitive load [6]–[10].

Research in software engineering is also starting to take advantage of biometric data to better understand what developers are going through in their work, measure their productivity, and to overall improve their productivity and wellness. With the recent advances, biometric sensors are becoming increasingly less invasive and are easier to accept and integrate into a developer's work without being bound to specific tasks, computers or locations. At the same time, the advances allow to capture more fine-grained biometric data in real-time, which offers new opportunities for more instantaneous support and feedback to developers. The vision is to integrate biometric sensing into a developer's work and use the data to ensure a developer's time is spent as productive as possible. In particular, biometrics might be used to measure aspects such as the flow and progress of individual developers or the difficulty they experience. These measures could then be used to provide instantaneous support, for instance, to avoid

interruptions at inopportune moments, detect difficult parts in the code, and to intervene before a developer creates a bug.

In this paper, we will present an overview of the use of biometric sensors in the context of software development in general and more specific findings from our initial studies that demonstrate the potential that biometric data can have to accurately and instantaneously measure perceived task difficulty, progress and interruptibility of a developer. This offers much promise to provide better developer support and improve individual productivity. At the same time there are still several challenges to overcome for this to become a reality and widely accepted by developers, such as privacy concerns or sensor limitations that we will discuss as well.

The paper is structured as follows. First, we provide background information on biometric data and the measures commonly used in research (Section II). Second, we present a general biometric sensing approach in software development (Section III) and discuss findings of initial research in the area, including ours (Section IV). Then, we discuss future opportunities and challenges (Section V) before we conclude.

II. Biometric Data

Psychophysiology explores the relation between psychological states and processes and their physiological reactions [11]. An increasing amount of research has shown that specific cognitive and emotional states, such as high or low cognitive load, or arousal and valence, can be correlated with biometric measures, such as electro-dermal activity or pupillary response [12]–[16]. These psychological states and processes are influenced by the person and the task that is being performed, and in turn affect the outcome. For instance, according to the cognitive load theory [17], cognitive load—the required mental effort to perform a task—is composed of intrinsic, extrinsic and germane load, including aspects such as the inherent task difficulty, the task format, as well as the person's age, experience and personality traits. The cognitive load experienced by a person during a task affects aspects, such as the likelihood of errors being created, a person's interruptibility and performance [18]–[20].

Similar to cognitive load, valence and arousal are concepts that can be influenced by various factors, such as task difficulty and personality traits [21]–[23], and in turn can influence the outcome, such as the perceived progress and emotions [24]–[26]. According to Russell's circumplex model, arousal and valence are the two cognitive dimensions of emotions [27], [28]. The arousal dimension indicates the amount of activation that is associated with an emotion, while the valence dimension is referring to the positive or negative character of the emotion. Various studies in the area of psychology have shown that biometrics can be used to determine the arousal and the valence dimension of emotions [14], [29], [30].

Based on the link between biometrics and psychological states, biometrics might allow us to better understand what a developer experiences during work and to accurately predict outcome aspects, such as the error rate. An overview of these concepts in the software development context is given in

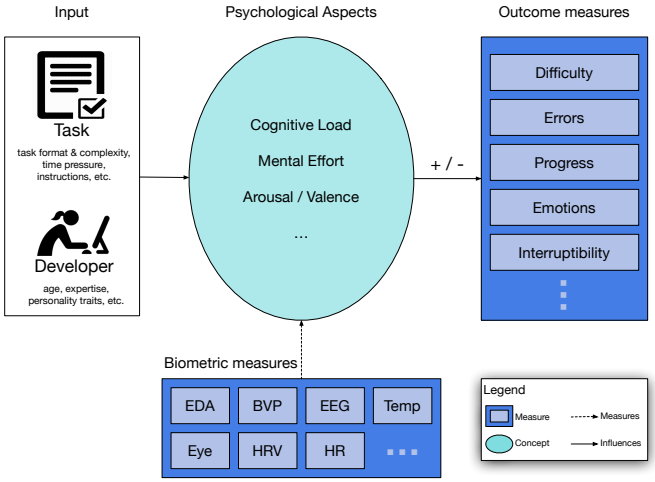

Fig. 1: Exemplary illustration of psychophysiological relations in a software development context.

TABLE I: Overview of some biometric measures and previously found links from literature.

Measure	Previously linked to
Eye-related	
Pupil size	cognitive & mental load [16], [31]; excitement [32];
Fixations	cognitive load [33]; valence [34]
Eye blinks	mental workload [35]; frustration, stress, anxiety [36], [37]
Brain-related	
Frequency bands (FBs)	mental workload [35]; valence, arousal [38], [39]; happiness and sadness [40]
Ratios of FBs	task engagement [41]; valence, arousal [42]
Skin-related	
Electro-dermal activity (EDA)	valence, arousal, engagement [15], [43]; frustration [36], [44]; stress and cognitive load [10]
Skin temperature	task difficulty [45]; valence, arousal [15]; boredom, engagement, anxiety [30];
Heart-related	
Heart rate (HR)	mental load & effort [13], [46]; valence, arousal [15], [38]; positive / negative affect [47]; happiness [48]
Heart rate variability (HRV)	mental effort [46]; task difficulty [49]; anxiety [50]; various emotional states [51]
Blood volume pulse (BVP)	cognitive load [52]; various emotions [53]; valence, arousal [15];
Breathing-related	
Respiratory rate	mental effort [46]; task difficulty [54]

Figure 1. Biometric measurements can roughly be divided into five different categories according to the origin of the measurements: eye-related, brain-related, skin-related, heart-related and breathing-related measurements. Table I provides an overview of some of these measures together with the psychological states and processes that previous research, mostly in psychology, has already linked them to.

Eye-related Measurements. Interesting eye-related features are the eye movement, the pupil size and the eye blinks. Eye movement can further be separated into fixations, when the eye gaze fixates on a specific, non-moving object, and saccades that refer to the moving of the gaze point from one object to another. Most of these features can be captured with an eye tracker that uses the reflection of infrared light from the eyes, but eye blinks can also be extracted from EEG data [55].

The average eye blink rate lies between 15 and 20 blinks per minute, but can increase significantly when a person is tired, experiences a lot of stress or is under time pressure [11]. Previous research has linked eye-related features to cognitive, mental and memory load [31], [35] as well as emotional aspects, such as valence or negative affect [32], [34].

Brain-related Measurements. The varying activity of neurons in the brain causes fluctuations in the voltage potential along the scalp that can be measured with an electroencephalogram (EEG) [11]. Research has identified a number of brain wave frequency bands from EEG data that are called alpha (α), beta (β), gamma (γ), delta (δ), and theta (θ). Each of these brain wave frequency bands has a specific frequency range and amplitude and exhibits more or less activity under certain circumstances. For instance, alpha waves can typically be observed when an individual is in a relaxed state, but the alpha waves either disappear or their amplitude decreases significantly as soon as the physical or mental activity increases [11]. Generally, research has linked these specific frequency bands and ratios thereof with mental workload, task engagement and emotions [35], [40], [41].

Skin-related Measurements. Common skin-related measurements are the skin temperature and the electro-dermal activity (EDA), formerly also known as galvanic skin response (GSR). EDA measures the electrical conductance of the skin. As an example, when an individual is aroused, the sweat glands in the skin will produce more sweat and the electrical conductance of the skin will therefore increase. The EDA signal consists of two parts: the slowly changing, low frequency, tonic part, and the fast adapting, high frequency, phasic part [56]. Commonly used features for the tonic part and the temperature signal are the mean value or the area under the curve (AUC); commonly used features for the phasic part are related to the peaks in the signal. EDA as well as skin temperature have previously been correlated with the general arousal level and also with specific emotions [15], [57], [58].

Heart-related Measurements. For heart-related measurements we focus on three different features: the heart rate (HR), the heart rate variability (HRV) and the blood volume pulse (BVP). The heart rate refers to the number of contractions of the heart each minute and the heart rate variability represents the variation in the time interval between two consecutive heart beats. The blood volume pulse measures the blood flow through specific parts of the body and may change when the sympathetic nervous system increases its activity, for instance because of stress [11]. In research, HR and HRV have been linked to mental and cognitive load, as well as stress levels and emotional states [13], [46], [51]. BVP has predominantly been correlated with various emotions [53]. Common features of these measurements are the mean heart rate, the mean and the standard deviation of the time between two heart beats and features that capture the peaks of the BVP signal.

Breathing-related Measurements. We only used one breathing-related measurement, the respiratory rate (RR). The respiratory rate refers to the number of breaths within a specific time period and under normal conditions it is in the range between 12 and 15 breaths per minute [59]. Research has used the respiratory rate to assess mental effort, task difficulty and task demand [46], [54]. Commonly used features are the mean respiration rate or the $\log_{10}$ variance of the respiration signal.

III. Biometric Sensing in Software Development

Biometric data has the potential to provide insights on what developers are experiencing in their work in real-time, for instance, when they are getting into the flow and are highly focused or when they are having difficulties and are getting frustrated. One of the early approaches in the software development domain mentioning biometric sensor technology is the Ginger2 environment by Torii and colleagues that included an eye-tracker and a skin resistance level sensor to empirically study developers [60]. In the context of software development, biometric sensors have been and are being used predominantly to gain a better understanding of developers' program comprehension either using eye-tracking [61]–[64] or by measuring brain activity [65], [66]. Recently, especially with the advances in sensor technology and the availability of more affordable biometric sensors, a few software engineering researchers have also looked at other aspects, for instance, using measures of cerebral blood flow, measures of sub-vocal utterances captured with electromyography (EMG), or EDA, eye-tracking and brain activity measures to asses task difficulty of small code snippets or programming tasks [67]–[69]. In a few cases, biometric measures have also been used for studies with software developers on longer and more realistic development tasks or even in the field [70]–[72].

To study the potential of biometric data for any particular aspect in the software development domain, there are a few general questions and steps that researchers have to address, such as which sensors to use, how to setup the study and how to analyze the data. In the following, we will discuss these three points and provide insights from our own experiences on the use of biometric sensors, before we will provide more detailed information on the use of biometric data to measure developers' perceived task difficulty, progress and interruptibility in Section IV.

A. Which Biometric Sensors to Use?

There is already a plentitude of devices available for a broad audience that contain some sort of biometric sensors, such as the Mio Fuse[1], the Microsoft Band[2] or the Apple Watch[3] that can track a person's heart rate. At this point, however, most of these devices do not yet support the granularity, sampling rate or the specific biometric features that are required and that were previously linked to psychological states and processes, such as high or low cognitive load. Therefore, more specialized biometric sensors are still needed for these kinds of studies, and it is not always easy to find a good set. Some of

[1] http://www.mioglobal.com
[2] http://www.microsoft.com/microsoft-band
[3] http://www.apple.com/watch

the questions to consider before choosing a set of sensor devices are: which biometric features have been shown to capture phenomena similar to the one of interest (*e.g.* in psychology research) and which set of sensors captures these; do the devices provide the necessary accuracy, granularity and sampling rate; do the sensor devices allow to conduct the research without obstructing/constraining the developer (too much); and is an API available to collect the biometric data from the devices as needed.

In our studies, we used seven different sensors so far: a Tobii TX300 eye tracker[4] or an Eyetribe eye tracker[5] for eye-related measures, an Affectiva[6] Q Sensor 2.0 (no longer available), or an Empatica E3 or E4 wristband[7] to record various skin- and heart-related measures, a Neurosky Mindband[8] to capture brain-related measures, and finally a SenseCore chest strap (also no longer available) to record various skin-, heart-, and breathing-related measurements. These specific sensors were chosen for three reasons: first, research in psychology has linked the features recorded with these sensors to outcome measures we are interested in, second, these sensors were minimally invasive for the measures they recorded, and third, these sensors were reasonably priced and affordable for a single developer, with the exception of the Tobii eye-tracker. The major challenges we encountered were due to a type of sensor ceasing to exist or being supported either because the company shifted focus or because the company actually closed down, and also due to the lack of mature APIs to collect certain data in real-time.

B. Setting up the Study

Most biometric features and sensors are sensitive to various variables, such as lighting and noise, the exact placing of the sensor devices, the time of day, or the weather [59], to name just a few. Therefore, when setting up a study with biometric sensors or when employing them in the field, it is important to think about how to best setup the study to examine the phenomena of interest. Relevant questions to address are: how do you ensure that the study conditions are the same or almost the same for all participants; should the study take place in a lab to better control the environmental variables, or in the field to examine whether the approach could also be used in more realistic settings; is it possible to capture the biometric data for longer periods or only for very short tasks; and how to best collect the outcome measure as well as a baseline of each biometric feature to normalize the captured biometric data and account for individual differences in the biometric data.

Our recent studies ranged from controlled lab experiments in which developers worked on short and small predefined code tasks to multiple days field studies with professional developers working in their usual environment on their usual tasks. Each study was designed to be minimally invasive in terms of time and impact on a participant's usual workflow to avoid biasing the results. Figure 2 depicts a participant in our field study on interruptibility [71] sitting in front of his work station while wearing an Empatica E3 wristband and a Neurosky Mindband. For this research, we ran a lab and a field study and found that biometric data can be used in both cases to predict interruptibility of developers, but that the predictions are more accurate in the lab. For the outcome measure, we used the Tablet to trigger interruptions and to collect user ratings on interruptibility and disruptiveness. In addition, we asked study participants to watch calming two-minute movies of fish swimming in a fish tank as well as to relax and not think of anything specific during that time period. We used the biometric data collected during the second minute of the movie as a baseline for each participant.

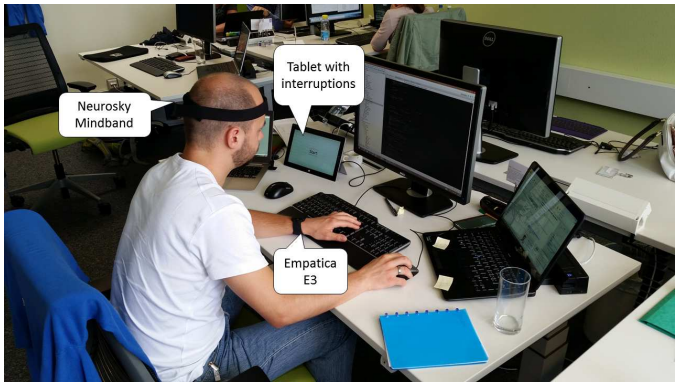

Fig. 2: Field study setup with a participant wearing a headband and a wristband. The tablet was used to trigger interruptions.

C. Data Analysis

Once the data is recorded, several steps to clean it and extract specific features of interest have to be performed. An overview of the general approach we followed in our studies to achieve this goal is presented in Figure 3.

Since biometric data is notoriously noisy, the first step is to clean the captured data. Depending on the kind of biometric data, different noise cleaning and filtering techniques have to be applied. For instance, for eye-tracking data the invalid data points that are labeled as such by the eye-tracker should be removed. For the EDA data an (exponential) smoothing filter can be applied to remove noise and the EDA signal's DC component should be subtracted to base it at $0\,\mu S$. Most of these noise cleaning techniques are described in literature, but also require a careful analysis of the captured data. During this cleaning process, it is also advantageous to segment the collected data as needed to reduce the amount of data that has to be processed for later analysis steps. In our previous study on predicting a developer's emotions and progress for example, we only used and processed ten-second time windows of biometric data collected just before each time we interrupted developers and asked them to rate their emotions and progress. For the segmentation it is important to make sure that the segments are independent from each other with respect to

[4]http://www.tobiipro.com/product-listing/tobii-pro-tx300/
[5]https://theeyetribe.com/
[6]http://www.affectiva.com/
[7]https://www.empatica.com
[8]http://neurosky.com/biosensors/eeg-sensor/

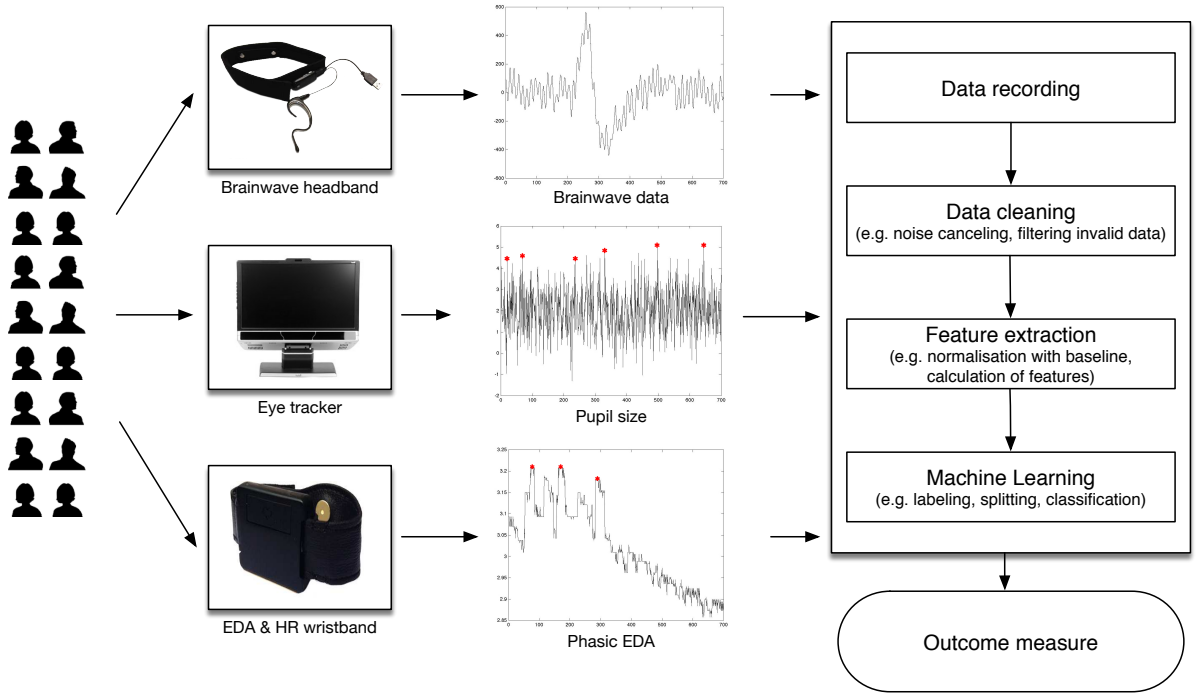

Fig. 3: Overview of the general approach to record and analyze biometric data.

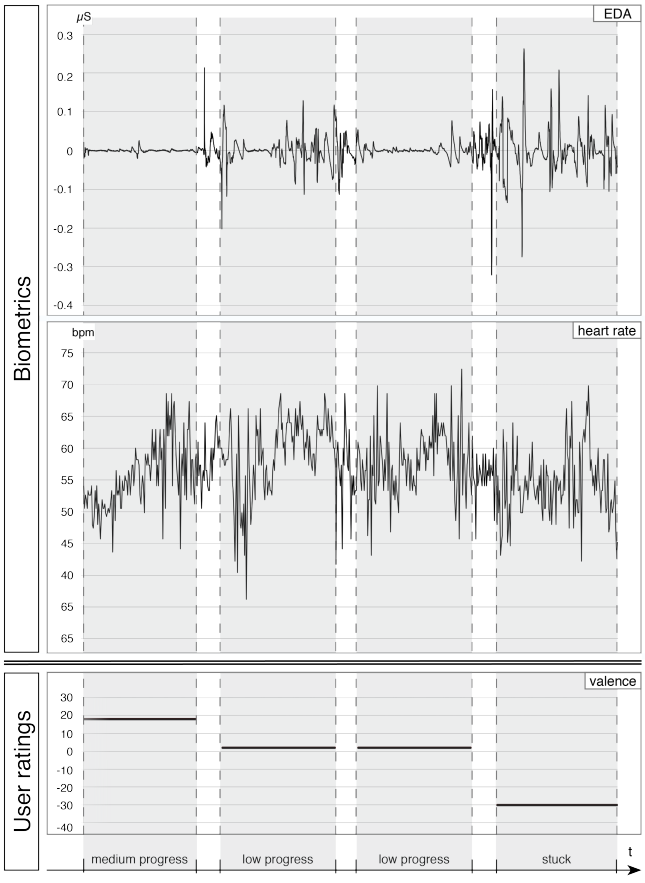

Fig. 4: Biometric data and the participant's perceived progress and valence ratings collected during one of our studies.

later analysis steps. Especially since biometric features, such as body temperature or the tonic part of the EDA signal, take some time to get back to a baseline or original value, the segmentation has to take this into account.

In many cases, the raw (and cleaned) data by itself is not meaningful and specific features have to be extracted. For instance, commonly extracted features from HRV data are the mean and the standard deviation of the time between two heart beats, and EEG data is commonly split up into five brain wave frequency bands ($\alpha - \theta$). Since biometric data is very individual, these features also have to be normalized by participant. In our previous studies, we used a biometric baseline collected during fish tank movies for the normalization.

Before feeding the extracted features into a machine learning classifier, the data needs to be labeled and split. For the data labeling, the outcome measure under analysis has to be assigned to the biometric data segments used for predicting the outcome measure. As an example from our research, Figure 4 presents biometric data—EDA and HR—that we collected during one of our studies in combination with the participant's valence and progress ratings that it was then labeled with [70]. Valence refers to the positive or negative character of an experienced emotion, the higher the rating the more positive the emotion. In the depicted data sample, a difference is visible between the EDA signal during the phases of medium progress and positive valence rating, and the phase in which the developer got stuck and perceived negative emotions. In a next step, the data has to be split into training and test data. Depending on the evaluation method (*e.g.* cross-validation or leave-one-out), different methods to perform the splitting have to be considered. In all cases, it is important to ensure the

training and test data set do not overlap or the splitting biases the predictions in any way.

IV. Sensing Difficulty, Progress and Interruptibility

One of the most common reasons for professional developers to have a productive workday is getting into the flow and making lots of progress without having many context-switches, interruptions or distractions [73]. While it was previously difficult or sometimes impossible to measure aspects such as the developer's perceived progress or interruptibility in real-time, the advances in biometric sensors might provide us the means to measure them. In the following, we will discuss the value and feasibility of using biometric sensors in the software development domain by focusing on three such aspects that we also explored in our own research: task difficulty, progress and interruptibility.

A. Sensing Task Difficulty

Knowing when and for which code or tasks developers experience difficulties might allow us to lower development and evolution cost, for instance by identifying quality concerns in the code early on and by intervening before developers create bugs. Research on manually detecting code quality concerns has shown that code reviews can help significantly to discover and improve code with quality concerns, including defects [74]–[76]. However, manual inspections require time and effort and can only be done after. Most research to automatically determine task and code difficulty as well as predicting defects has predominantly focused on the use of various code metrics, such as complexity metrics, size metrics, or code churn [3], [77]–[80]. These metrics, however, can mostly only be computed after a developer finished a change task and also do not take into account the individual differences that exist between developers. A step towards more individual data of developers was taken by Lee and colleagues who defined micro interaction patterns for predicting defects [81].

Since research in psychology has shown that certain biometric features are linked to a person's mental effort for working on a task, biometric data has a great potential to help us assess the difficulty a developer is experiencing when working on a change task and when the developer might be close to creating an error in the code. Some preliminary work in software engineering has looked at the use of biometric measurements to determine task difficulty. For instance, Nakagawa *et al.* [67] measured cerebral blood flow (CBF) to distinguish between two difficulty levels while developers were performing code comprehension tasks. Similarly, Parnin [68] investigated the use of electromyography (EMG) to measure sub-vocal utterances and found that these measurements might be used to assess programming task difficulty.

In our own work with A. Begel, S. Yigit-Elliott and M. Züger, we conducted a controlled lab experiment with 15 professional software developers to examine the feasibility of using biometric sensors to assess the difficulty a developer experiences working on small code comprehension tasks [69]. Each participant was asked to perform eight short code comprehension tasks while sitting in front of a computer with an eye-tracker and wearing an EDA and an EEG sensor. For each task participants were asked to read a small C# code snippet and then answer a multiple choice question. To ensure varying levels of task difficulty, we altered tasks in several ways, such as the use of obfuscated variable names instead of mnemonic ones, randomly-ordered field assignments or the use of loops with various levels of complexity. With these tasks and variations we wanted to stress participants' cognitive abilities, such as the working memory, their math and logic skills and their ability in spatial relations. For assessing perceived task difficulty, we asked all participants to rank and rate the tasks according to their difficulty.

Findings. We trained a Naïve Bayes classifier and found that we are able to use the collected biometric data to predict the perceived task difficulty for a new developer that we did not train on with 65.0% precision and 64.6% recall using all three sensors. The precision and recall went up to 84.4% and 69.8% respectively, when we predicted for a new task and trained the classifier on other tasks for each developer. So while the classifiers can be used to predict difficulty even for people they were never trained on, they can be a lot better in predicting when they are trained on the individual they are used for, since biometric data varies a lot across people. Our analysis also showed that, while subsets of the biometric sensors also lead to good results, a combination of all three sensors, eye-tracking, EDA and EEG, performed best. Finally, an analysis on sliding time windows of the biometric data also showed that even short windows of only a couple of seconds can achieve high precision and recall, illustrating the potential of measuring task difficulty in real-time [69].

B. Sensing Progress and Emotions

Developers feel particularly productive in a workday when they get into the flow, making lots of progress [73]. If we are able to measure a developer's progress and flow in real-time, we would, for example, know when a developer is stuck and she might need help or even just a break, or we could use the information on a developer being in the flow to indicate to coworkers not to interrupt and avoid costly interruptions. Similarly, knowing more about a developer's emotions might allow us to help them when they are frustrated and would benefit from a break or the help of a coworker.

Several empirical studies have investigated the kind of emotions developers experience, their change over time and their correlation with developers' productivity, also, for instance, by inducing moods and measuring the impact on developers' performance [25], [26], [82]–[84]. Khan *et al.* also investigated whether keyboard and mouse input could be used to measure a developer's mood, but no generic correlation across all study participants was found [85]. Concerning the classification of progress, only very few studies have been conducted in the field of software engineering. Carter *et al.*, for instance,

mined IDE interaction logs to automatically determine when a developer is stuck and cannot make any progress [86].

In one of our studies, we examined the use of biometric data to assess a developer's progress and emotions, especially since research in psychology has already linked progress and emotions [70]. For this study, we had 17 participants (6 professional software developers, 11 Computer Science PhD students) work on two change tasks on open source systems for 30 minutes each in a quiet lab while again sitting in front of an eye-tracker and wearing a wristband to track EDA and heart-related measures as well as an EEG sensor. We used experience sampling, interrupting participants approximately every five minutes and asking them to rate their progress on a 5-point Likert scale and their emotions according to Russell's 2-dimensional Circumplex model on a valence and an arousal scale each from -200 to +200 [27]. As a baseline for our emotion-related measures, we also showed participants sets of pictures that are known to induce positive and negative emotions [87]. We used ten second time windows of biometric data just before each interruption and labeled these as either positive or negative emotion instances by using the baseline's ratings to split valence ratings during a change task as either positive or negative instances. For the progress ratings, we labeled the biometric data segments as high progress for ratings of 4 and 5 and low for 1 and 2, while we also removed neutral instances. We then used the collected and labeled 10 second time windows of biometric data to train and test a decision tree (J48) classifier using a leave-one-out method.

Findings. Using a decision tree classifier we were able to correctly classify cases as low or high progress in 67.7% of all cases (improving upon a naive classifier by 32.9% and a random one by 35.4%), and as negative or positive emotions in 71.4% of all cases (improving upon a naive classifier by 18.8% and a random one by 42.7%). The analysis also showed that there is a big variance between individuals and in how accurately one can predict progress and emotions using biometric data ranging from as little as 30.0% for one participant to getting all cases right (100%) for another one. The biometric measures with the most predictive power for progress was thereby again a combination of all three biometric sensors: EDA tonic signal, skin temperature, brainwave frequency bands and pupil size. Overall, the results of this study provide further evidence on the value that biometric data can have to measure aspects of a developer during work and that it also works for longer and more realistic change tasks [70].

C. Sensing Interruptibility

Interruptions were one of the most commonly named reasons for decreasing developer productivity [73]. Studies have actually shown that interruptions at inopportune moments will not only slow down a developer's work significantly, they will also lead to negative emotions and more errors in the code [88], [89]. An automatic measure of interruptibility could thus significantly increase developers' productivity, for instance, by reducing in-person interruptions through visual cues to coworkers or by postponing computer-based interruptions, such as instant messages.

Much research on assessing interruptibility investigated the simulation and use of context-aware sensors, such as audio or video streams, keyboard or mouse input, active window information or information on task characteristics. For instance, Hudson *et al.* were able to classify interruptibility into two states with 78% accuracy by manually coding video and audio streams based on features, such as the phone being on the hook or people speaking [90]. Fogarty *et al.* also simulated sensors—manually coding mouse and keyboard interactions—and were able to predict two states of interruptibility based on the interruption lag with 72% accuracy [91]. More recently, researchers started to explore the use of context-aware sensors with no need for manual coding. Tani *et al.* used pressure sensors to measure the force applied when typing on the keyboard and using the mouse. They were able to classify interruptibility into two states with an accuracy of around 70% [92]. Ho *et al.* used accelerometers to identify physical activity transitions and found that interruptions at activity transitions are perceived better than those at random times [93].

Fewer research looked at the use of biometric sensors to assess interruptibility. Mathan *et al.* used EEG data to classify interruptibility during a US army urban combat training mission [94]. Bailey *et al.* focused on eye-related measurements (pupil size) to infer the mental workload of study participants working on a goal-directed task. They found that a user's mental workload decreases at (sub)task boundaries suggesting that interruptions are best at these boundaries when the workload is low and fewer resources are needed to resume the task [95]. Finally, Chen *et al.* used heart rate variability (HRV) and an electromyogram (EMG) and found a strong and significant correlation between the biometric measurements and the users' self-reported interruptibility during a variety of short and simple tasks [96].

In our work, we extended these studies by exploring the use of biometric sensors in real-world working contexts of software developers [71]. In particular, we conducted two studies, a lab and a field study. In the lab study, we had eight graduate students work on three realistic change tasks on JHotDraw for a total of 60 minutes per student. For the field study, we recruited and visited ten professional software developers from four different companies and had them work on their own tasks in their real-world office environments for two hours each. Participants were told to work as usual without restriction on their activities. For both studies, participants were asked to wear a headband to record EEG and eye blink data and a wristband to record skin- and heart-related measures. During both studies, we triggered interruptions by playing a sound and changing the display on a tablet that we placed next to the developer's monitor(s) (see Figure 2). These interruptions were negotiated interruptions, i.e. participants could decide for themselves when to address them, and were triggered at random time intervals that were

between one and eleven minutes long. For each interruption we asked participants to perform a mental arithmetic exercise and to rate their interruptibility at the time of the notification and the perceived disturbance of the interruption each on a 5-point Likert scale. For our analysis we used time windows of biometric data ranging from ten seconds to three minutes ending with the triggered interruption. We labeled the time windows with user ratings, once categorizing the 5-point scale into two states (interruptible or not) and once by keeping the fine-grained five state classification. For the machine learning we used Naïve Bayes since it outperformed Decision Trees and Support Vector Machine approaches with a ten-fold cross-validation.

Findings. For the lab as well as the field study we were able to use the biometric data to classify a developer's interruptibility with high accuracy into two states (lab: 91.5%, field: 78.6%) and into five states (lab: 43.9%, field: 32.5%). In all cases except for the five-state classification in the field, the trained classifier significantly outperforms a simple majority classifier. The fact that the lab study results are better than the field study ones and that the five-state classification for the field study did not significantly outperform the majority classifier, hints to the effect that external influences and more noise can have on such sensors in the field. The analysis also showed that shorter time windows of 10 seconds work generally better. Overall, these results illustrate the high potential of using biometric data as real-time indicators for interruptibility, and that they can be transferred to a real-world environment [71].

V. Opportunities and Challenges

Overall, results of recent research studies demonstrate the high potential that biometric data has to measure certain aspects of a developer's work in real-time and with low-cost, off-the-shelf biometric sensors. In particular, the results show that biometric data, even captured in the field and only using short time windows, allows to accurately predict the difficulty developers experience during their work as well as their progress and interruptibility.

At some point in the future, we should be able to collect and leverage biometric data for each developer just the way that we are currently collecting and leveraging data on the artifacts that developers produce. However, instead of only calculating metrics on already completed change task data, we would then be able to know more about each individual developer while she is working on a change task. This will open up tremendous new opportunities for providing better developer support and boosting the productivity of each individual developer. Especially with the fast advances and the pervasiveness of sensor technology, we might soon be able to collect valuable biometric data for a broad audience of software developers and take advantage of some of these opportunities in practice. At the same time, the sensitivity of the data and aspects of the technology also pose new challenges, such as privacy and ethical concerns.

A. Opportunities

The findings of our research and other studies open up new opportunities to better support every individual developer in real-time to boost their productivity and general well-being.

Measuring and Supporting Developers in Real-Time. Traditional approaches to assess task difficulty, code quality and a developer's productivity predominantly used various kinds of code and process metrics defined on a developer's output. In most cases, these metrics are only available after a developer completed a task and thus only allow for a post-hoc analysis. Biometric data on the other hand has the potential for real-time measures and our study results have provided initial evidence on the accuracy and feasibility of such measures, especially given the short time windows of biometric data required.

Real-time measures on what a developer is experiencing, such as the difficulty of the task at hand or how focused the developer is at a given point in time also enable new possibilities for supporting developers while they are working. For example, a real-time measure of the difficulty a developer experiences while working on a specific code element might be used to automatically detect quality concerns in the code or when a developer is likely to create a bug in the code. This in turn would allow us to automatically prioritize code reviews and focus attention on the parts of the code that need it the most, and it might even enable us to prevent developers from creating bugs or committing them to a repository. Another example of the high potential of such biometric measures are real-time measures on a developer's interruptibility and flow. By knowing when a developer is highly focused and in the flow, we are able to provide awareness to coworkers and thus decrease costly interruptions. At the same time, we could use such a measure in combination with persuasive technologies such as self-monitoring and goal-setting which could foster an increase in productivity, similar to the way that activity tracking devices help to increase and maintain physical activity over extended periods of time [97].

Measuring Individual Developers. Another drawback of traditional metrics is that they are mostly just based on artifacts and do not take into account the individual differences that exist between developers, such as their experience and expertise, but also their mood or general well-being that might influence their output and productivity. Biometric measures shift the focus to the individual developers. While this might require training such measures or classifiers for each individual developer, this will allow for better performance and more accurate predictions as our studies have shown.

Each person has times during the day or even days when they are more or less focused and productive. In psychology, some studies distinguish, for example, between morning and evening people [98]. Measures that allow us to capture the cognitive and emotional states of each individual, will allow us to provide more tailored and valuable support. For instance, by knowing when a developer is most productive during the day, we might be able to optimize the work day and schedule the most demanding tasks for these times or avoid costly

interruptions. We would also be able to provide more tailored and in-time recommendations, such as suggesting to talk to a coworker for help or taking a break when a developer is stuck and frustrated.

Boosting Productivity and Wellness. Overall, the insights that biometric data can provide us on a developer's work have immense potential to boost productivity. Preventing bugs, reducing quality concerns, avoiding interruptions at inopportune moments, providing a retrospective analysis, and automatically scheduling a productive day for developers are just some of the opportunities that biometric data might enable. All of these can significantly improve a developer's work flow and productivity as well as reduce software evolution cost. Once we start taking advantage of biometric data, there are plenty of opportunities to amplify the human smarts and ingenuity in the development process, and also extend it to a broader range of stakeholders in the process, such as managers, testers or operators.

Biometric data might also be used to assess and foster the overall wellbeing of developers and in turn again their productivity. Several study results have already highlighted the correlation between progress and positive emotions, such as happiness. On a larger scale, it is also known that employees that are happy with their work and company have less sick days and are less likely to quit their job, which in turn increases the overall productivity of the company. With an increased awareness of the wellbeing of the developers on a team or in the company, managers might be able to react more quickly and provide a better work environment for their developers.

Pervasiveness of Smart Wearable Devices. The term smart wearable devices refers to electronic devices that are integrated in accessories or clothes and provide their users with some kind of real-time feedback about the user's activities or physiological features. The market for smart wearable devices is growing at an immense speed. From a market value of 600 million US$ in 2013, the market has increased to around 4 billion US$ in 2014 and will, according to forecasts, reach 30 billion US$ in 2020 [99]. Wearable devices come in many different forms, such as smart glasses like the Google Glass[9], smart watches like the Apple Watch[10] or fitness tracker like the Fitbit[11].

The biometric sensors that are integrated in some of these devices are becoming more and more sophisticated and will soon support the capturing of biometric data that might be fine-grained and accurate enough for some of the presented scenarios. With the fast growing market and pervasiveness of smart wearable devices, an increasing amount of people will be wearing such devices. This will provide the opportunity to capture and collect biometric data for many developers on a daily basis and allow for the fine-tuning of classifiers on aspects such as task difficulty or progress which will further increase the potential the data has.

[9]https://www.google.com/glass/start/

[10]http://www.apple.com/watch

[11]https://www.fitbit.com/

B. Challenges

There are also several challenges with the use of biometric data, some of which we address in the following.

Sensor Limitations. To collect the relevant biometric data, you need sensors that are able to capture it and people that are willing to wear these sensors. Especially to collect data over longer periods of time, sensors need to be minimally invasive. In most cases though, there is a trade-off between invasiveness and the granularity and accuracy of the data that can be captured by a sensor. For instance, while some of our study participants agreed to wear a headband during a two hour study that allowed us to collect EEG data, most people would not wear this over several work days or even a whole work day and thus EEG data might not always be available. While there are new and less invasive sensing devices coming out frequently, for instance the Apple Watch, it will still take some time until certain biometric features will be accessible over longer periods of time without disturbing the wearer. Furthermore, since several of these devices are still relatively new and immature, including their data transfer and the provided APIs, there are still quite a few challenges and obstacles to overcome to collect the relevant data. However, the fast growth of the market and technology should improve this situation soon.

Privacy & Ethical Concerns. The sensitivity and amount of the captured biometric data raises several privacy and ethical concerns. Using such data, for instance, to assess a developer's productivity or other work-related aspects might lead to a big brother effect and raise several red flags in developers. For our studies, we approached these concerns by ensuring to store the data in an anonymized way and by only making it accessible to the individual participant and the researchers. However, to provide the benefits to a broader audience of professional developers and possibly also their managers, future research has to look into means to ensure privacy and security of the biometric data. In particular, research has to examine which abstractions and aggregations of the data are most valuable while still ensuring enough privacy for the individual.

Recruiting Study Participants. Due to the sensitivity of the collected data, the invasiveness of the sensors, and general privacy and time concerns of software developers, finding participants for studies with biometric sensors is tedious and time-consuming. This situation is further aggravated by the fact that study participants can not always see an immediate benefit from participating. Especially for longer, several-day studies, it can be very challenging to find professional developers. This problem can be mitigated to some extent by choosing sensors that are as little invasive as possible, letting participants access their own biometric data and continuously motivating them to participate.

Noisy Data. Research has shown that biometric measures can be affected by the weather, the time of day, or personality traits [59]. Thus the collected biometric data might contain a lot of noise and be affected by many other aspects than just the outcome measures, including individual differences.

To minimize some of these effects, the environment in which a study will be conducted has to be chosen carefully, baselines should be established for each participant and normalization techniques should be applied. The results of our field studies provide initial evidence that biometric sensors can be used in noisy environments to make accurate predictions [71], [72]. With the new advances in sensor technology and also by training classifiers on bigger data sets of an individual, it might be possible to minimize the effects of the noise even further and provide more fine-grained classifications.

Choosing the Best Biometric Features. Research in psychology has analyzed various biometric features and found correlations to a lot of psychological aspects in some studies but none in others. This makes it difficult to know which features are best and which ones should be chosen for a machine learning classifier in a specific scenario. More research is therefore needed to establish which features are most promising in which contexts.

Analyzing Big Data. Typically, biometric sensors used for these kinds of studies have a relatively high sampling rate and generate big amounts of data even for smaller studies. For instance, over the course of our studies, we captured biometric data consisting of close to 140 million data points. This big amount of data leads to challenges in handling the data, cleaning the noise, normalizing it and extracting the necessary features. Some of these issues can be tackled by having dedicated machines, and where possible, parallelizing the algorithms. By optimizing the data collection step *e.g.* by focusing on a subset of features, reducing the sampling rate, or limiting the time windows, and with the advances in technology in general, it should be possible to further speed up the analysis process significantly in the near future.

VI. Conclusion

Biometric data has the potential to reveal a lot about a developer's cognitive and emotional states in real-time. Initial results from several studies confirm this hypothesis and show that biometric data can actually be used to accurately predict certain aspects of a developer in real-time, such as the experienced task difficulty, the perceived progress and the developer's interruptibility. This offers much promise for improving developer support and boosting developer productivity overall, for instance, by automatically and instantaneously detecting code quality concerns or by reducing costly interruptions.

Our vision is to leverage biometric data on developers during their work just the way we are currently leveraging more traditional metrics, and then be able to provide developers with better and more individually tailored support. Especially with the fast advances in sensor and data analysis technology, we might soon all be wearing smart wearable devices with biometric sensors integrated that will already be accurate enough to provide some of this support. Given the sensitivity and the amount of biometric data collected per individual, there are however still several challenges to be addressed, in particular privacy concerns of the data and challenges for conducting research in the area.

References

[1] M. Andreessen, "Why software is eating the world," The Wall Street Journal, August 20, 2011.

[2] "www.idc.com/getdoc.jsp?containerId=244709."

[3] N. Nagappan and T. Ball, "Use of relative code churn measures to predict system defect density," in *Proceedings of International Conference on Software Engineering*, 2005.

[4] H. Zhang, X. Zhang, and M. Gu, "Predicting defective software components from code complexity measures," in *Proc. of the Pacific Rim Intern. Symp. on Dependable Computing*, 2007.

[5] W. S. Humphrey, *Introduction to the Personal Software Process*, 1st ed. Addison-Wesley Professional, 1996.

[6] A. F. Kramer, "Physiological metrics of mental workload: A review of recent progress," *Multiple-task performance*, pp. 279–328, 1991.

[7] A. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with EEG pattern recognition methods," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 40, no. 1, pp. 79–91, 1998.

[8] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, space, and environmental medicine*, vol. 78, pp. B231–B244, 2007.

[9] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo, "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, 2012, pp. 420–423.

[10] C. Setz, B. Arnrich, J. Schumm, R. L. Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *Trans. on Information Technology in Biomedicine*, vol. 14, no. 2, 2010.

[11] J. L. Andreassi, *Psychophysiology: Human Behavior & Physiological Response*. Lawrence Erlbaum Associates, 2007.

[12] G. F. Wilson, "An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures," *International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, 2002.

[13] P. Richter, T. Wagner, R. Heger, and G. Weise, "Psychophysiological analysis of mental load during driving on rural roads - a quasi-experimental field study," *Ergonomics*, vol. 41, no. 5, 1998.

[14] M. Murugappan, M. Rizon, R. Nagarajan, and S. Yaacob, "EEG feature extraction for classifying emotions using fcm and fkm," in *Proceedings of the 7th WSEAS International Conference on Applied Computer and Applied Computational Science*, 2008, pp. 299–304.

[15] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," *Affective Dialogue Systems Lecture Notes in Computer Science*, vol. 3068, pp. 36–48, 2004.

[16] S. T. Iqbal, X. S. Zheng, and B. P. Bailey, "Task-evoked pupillary response to mental workload in human-computer interaction," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 2004, pp. 1477–1480.

[17] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*. Springer, 2011.

[18] A. J. Ko and B. A. Myers, "A framework and methodology for studying the causes of software errors in programming systems," *Journal of Visual Languages & Computing*, vol. 16, no. 1, pp. 41–84, 2005.

[19] S. T. Iqbal and B. P. Bailey, "Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1489–1492.

[20] P. Ayres, "Systematic mathematical errors and cognitive load," in *Contemporary Educational Psychology*, 2001.

[21] J. Basch and C. D. Fisher, "Affective events - emotions matrix: A classification of work events and associated emotions," in *Proceedings of the First Conference on Emotions in Organizational Life*, 1998.

[22] R. Plutchik and H. R. Conte, *Circumplex Models of Personality and Emotions*. American Psychological Association, 1997.

[23] J. Hutt and G. Weidner, "The effects of task demand and decision latitude on cardiovascular reactivity to stress," *Behavioral Medicine*, 1993.

[24] A. P. Brief and H. M. Weiss, "Organizational behavior: affect in the workplace." *Annual Review of Psychology*, vol. 53, pp. 279–307, 2002.

[25] D. Graziotin, X. Wang, and P. Abrahamsson, "Are happy developers more productive? the correlation of affective states of software developers ane their-self-assessed productivity," in *Proceedings of the 14th International Conference on Product-Focused Software Process Improvement*, 2013, pp. 50–64.

[26] I. A. Khan, W.-P. Brinkman, and R. M. Hierons, "Do moods affect programmers' debug performance?" *Cognition, Technology & Work*, vol. 13, no. 4, pp. 245–258, 2011.

[27] J. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[28] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological Review*, vol. 110, no. 1, pp. 145–172, Jan 2003.

[29] M. Murugappan, R. Nagarajan, and S. Yaacob, "Modified energy based time-frequency features for classifying human emotions using eeg," in *Proceedings of the International Conference on Man-Machine Systems*, 2009.

[30] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," in *Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*, 2008, pp. 13–17. [Online]. Available: http://doi.acm.org/10.1145/1457199.1457203

[31] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psychophysiological measures for assessing cognitive load," in *Proceedings of the 12th International Conference on Ubiquitous Computing*, 2010, pp. 301–310.

[32] K. Muldner, W. Burleson, and K. VanLehn, ""Yes!": using tutor and sensor data to predict moments of delight during instructional activities," in *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization*, 2010, pp. 159–170.

[33] C. S. Ikehara and M. E. Crosby, "Assessing cognitive load with physiological sensors," in *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005, p. 295a.

[34] E. Carniglia, M. Caputi, V. Manfredi, D. Zambarbieri, and E. Pessa, "The influence of emotional picture thematic content on exploratory eye movements," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–9, 2012.

[35] J. B. Brookings, G. F. Wilson, and C. R. Swain, "Psychophysiological responses to changes in workload during simulated air traffic control," *Biological Psychology: Psychophysiology of Workload*, vol. 42, no. 3, pp. 361 – 377, 1996.

[36] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, 2007.

[37] D. G. Doehring, "The relation between manifest anxiety and rate of eyeblink in a stress situation," Central Institute for the Deaf, St Louis, MO, Tech. Rep., 1957.

[38] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch, "Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music." *Psychophysiology*, vol. 44, no. 2, pp. 293–304, 2007.

[39] B. Reuderink, C. Mühl, and M. Poel, "Valence, arousal and dominance in the eeg during game play," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 6, no. 1, pp. 45–62, 2013.

[40] M. Li and B.-L. Lu, "Emotion classification based on gamma-band EEG." *Conference Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1323–1326, 2009.

[41] A. F. Kramer, "Physiological metrics of mental workload: A review of recent progress," *Multiple-task Performance*, pp. 279–328, 1991.

[42] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[43] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "Affectaura: an intelligent system for emotional memory," in *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*, 2012, pp. 849–858.

[44] G. L. Freeman, "A method of inducing frustration in human subjects and its influence upon palmar skin resistance," *The American Journal of Psychology*, vol. 53, no. 1, pp. pp. 117–120, 1940.

[45] L. Anthony, P. Carrington, P. Chu, C. Kidd, J. Lai, and A. Sears, "Gesture dynamics: Features sensitive to task difficulty and correlated with physiological sensors," *Stress*, vol. 1418, no. 360, 2011.

[46] J. Veltman and A. W. Gaillard, "Physiological workload reactions to increasing levels of task difficulty." *Ergonomics*, vol. 41, no. 5, pp. 656–669, 1998.

[47] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen, "Correlation between heart rate, electrodermal activity and player experience in first-person shooter games," in *Proceedings of the 5th Symposium on Video Games*, 2010, pp. 49–54.

[48] A. Steptoe, J. Wardle, and M. Marmot, "Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes." *Proceedings of the National Academy of Sciences*, vol. 102, no. 18, pp. 6508–6512, 2005.

[49] G. F. Walter and S. W. Porges, "Heart rate and respiratory responses as a function of task difficulty: The use of discriminant analysis in the selection of psychologically sensitive physiological responses," *Psychophysiology*, vol. 13, no. 6, 1976.

[50] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system - towards implicit human-robot collaboration," *Robotica*, vol. 22, no. 1, pp. 85–95, 2004.

[51] R. McCraty and D. Tomasino, *Stress in Health and Deseases*. Wiley-VCH, 2006, ch. Emotional Stress, Positive Emotions, and Psychophysiological Coherence.

[52] E. Peper, R. Harvey, I.-M. Lin, H. Tylova, and D. Moss, "Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony?" *Biofeedback*, vol. 35, no. 2, 2007.

[53] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[54] N. A. Kuznetsov, K. D. Shockley, M. J. Richardson, and M. A. Riley, "Effect of precision aiming on respiration and postural-respiratory synergy," *Neuroscience letters*, vol. 502, no. 1, pp. 13–17, 2011.

[55] P. Manoilov, "Eye-blinking artefacts analysis," in *Proceedings of the International Conference on Computer Systems and Technologies*, 2007, p. 52.

[56] S. Schmidt and H. Walach, "Electrodermal activity (EDA) - state-of-the-art measurements and techniques for parapsychological purposes," *Journal of Parapsychology*, vol. 64, no. 2, p. 139, June 2000.

[57] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions." *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[58] W. Boucsein, *Electrodermal Activity*. Springer, 2012.

[59] J. Cacioppo, L. G. Tassinary, and G. G. Berntson, *The Handbook of Psychophysiology*. Cambridge, 2007.

[60] K. Torii, K.-i. Matsumoto, K. Nakakoji, Y. Takada, S. Takada, and K. Shima, "Ginger2: An environment for computer-aided empirical software engineering," *Software Engineering, IEEE Transactions on*, vol. 25, no. 4, pp. 474–492, 1999.

[61] R. Bednarik and M. Tukiainen, "An eye-tracking methodology for characterizing program comprehension processes," in *Proceedings of the Symposium on Eye Tracking Research & Applications*, 2006, pp. 125–132.

[62] M. Crosby and J. Stelovsky, "How do we read algorithms? A case study," *Computer*, vol. 23, no. 1, pp. 25–35, 1990.

[63] P. Rodeghero, C. McMillan, P. W. McBurney, N. Bosch, and S. D'Mello, "Improving automated source code summarization via an eye-tracking study of programmers," in *Proceedings of International Conference on Software Engineering*, 2014.

[64] K. Kevic, B. M. Walters, T. R. Shaffer, B. Sharif, D. C. Shepherd, and T. Fritz, "Tracing software developers' eyes and interactions for change tasks," in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. ACM, 2015, pp. 202–213.

[65] Y. Ikutani and H. Uwano, "Brain activity measurement during program comprehension with NIRS," in *Proceedings of the International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 2014, pp. 1–6.

[66] J. Siegmund, C. Kästner, S. Apel, C. Parnin, A. Bethmann, T. Leich, G. Saake, and A. Brechmann, "Understanding understanding source code with functional magnetic resonance imaging," in *Proceedings of the 36th International Conference on Software Engineering*, 2014, pp. 378–389.

[67] T. Nakagawa, Y. Kamei, H. Uwano, A. Monden, K. Matsumoto, and D. M. German, "Quantifying programmers' mental workload during program comprehension based on cerebral blood flow measurement: A controlled experiment," in *Companion Proceedings of International Conference on Software Engineering*, 2014.

[68] C. Parnin, "Subvocalization - toward hearing the inner thoughts of developers," in *Proceedings of the International Conference on Program Comprehension*, June 2011, pp. 197–200.

[69] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using psycho-physiological measures to assess task difficulty in software development," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 402–413. [Online]. Available: http://doi.acm.org/10.1145/2568225.2568266

[70] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *Proceedings of the 37th International Conference on Software Engineering - Volume 1*, ser. ICSE '15. Piscataway, NJ, USA: IEEE Press, 2015, pp. 688–699. [Online]. Available: http://dl.acm.org/citation.cfm?id=2818754.2818838

[71] M. Züger and T. Fritz, "Interruptibility of software developers and its prediction using psycho-physiological sensors," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, 2015, pp. 2981–2990. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702593

[72] S. C. Müller and T. Fritz, "Using (bio)metrics to predict code quality online," in *Proceedings of ICSE'16*, 2016, to appear.

[73] A. N. Meyer, T. Fritz, G. C. Murphy, and T. Zimmermann, "Software developers' perceptions of productivity," in *Proceedings of the International Symposium on Foundations of Software Engineering*, 2014, pp. 19–29.

[74] A. Bacchelli and C. Bird, "Expectations, outcomes, and challenges of modern code review," in *Proceedings of the 35th International Confernce on Software Engineering*, 2013, pp. 712–721.

[75] A. Bosu, M. Greiler, and C. Bird, "Characteristics of useful code reviews: An empirical study at microsoft," in *Proceedings of the International Conference on Mining Software Repositories*, 2015, pp. 146–156.

[76] P. C. Rigby, D. M. German, and M.-A. Storey, "Open source software peer review practices: A case study of the apache server," in *Proceedings of International Confernce on Software Engineering*, 2008.

[77] N. Kasto and J. Whalley, "Measuring the difficulty of code comprehension tasks using software metrics," in *Proceedings of the Australasian Computing Education Conference*, 2013, pp. 59–65.

[78] B. Katzmarski and R. Koschke, "Program complexity metrics and programmer opinions," in *Proceedings of the International Conference on Program Comprehension*, 2012, pp. 17–26.

[79] J. Feigenspan, S. Apel, J. Liebig, and C. Kastner, "Exploring software measures to assess program comprehension," in *International Symposium on Empirical Software Engineering and Measurement*, 2011.

[80] D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson, "Using the support vector machine as a classification method for software defect prediction with static code metrics," in *Engineering Applications of Neural Networks*, D. Palmer-Brown, C. Draganova, E. Pimenidis, and H. Mouratidis, Eds. Springer Berlin Heidelberg, 2009.

[81] T. Lee, J. Nam, D. Han, S. Kim, and H. P. In, "Micro interaction metrics for defect prediction," in *Proceedings of the 19th Symposium and the 13th European Conference on Foundations of Software Engineering*, 2011, pp. 311–321.

[82] T. Shaw, "The emotions of systems developers: An empirical study of affective events theory," in *Proceedings of the Conference on Computer Personnel Research: Careers, Culture, and Ethics in a Networked Environment*, 2004, pp. 124–126.

[83] D. Graziotin, X. Wang, and P. Abrahamsson, "Happy software developers solve problems better: psychological measurements in empirical software engineering." *PeerJ*, vol. 2:e289, 2014.

[84] M. Wrobel, "Emotions in the software development process," in *Proceedings of the International Conference on Human System Interaction (HSI)*, 2013, pp. 518–523.

[85] I. A. Khan, W.-P. Brinkman, and R. Hierons, "Towards estimating computer users' mood from interaction behaviour with keyboard and mouse," *Frontiers of Computer Science*, vol. 7, no. 6, pp. 943–954, 2013.

[86] J. Carter and P. Dewan, "Design, implementation, and evaluation of an approach for determining when programmers are having difficulty," in *Proceedings of the Conference on Supporting Group Work*, 2010, pp. 215–224.

[87] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance." *Behavior research methods*, vol. 43, 2011.

[88] B. P. Bailey and J. A. Konstan, "On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state," *Computers in Human Behavior*, vol. 22, no. 4, pp. 685 – 708, 2006.

[89] M. Czerwinski, E. Horvitz, and S. Wilhite, "A diary study of task switching and interruptions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 175–182.

[90] S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee, and J. Yang, "Predicting human interruptibility with sensors: A wizard of oz feasibility study," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2003, pp. 257–264.

[91] J. Fogarty, A. J. Ko, H. H. Aung, E. Golden, K. P. Tang, and S. E. Hudson, "Examining task engagement in sensor-based statistical models of human interruptibility," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2005, pp. 331–340.

[92] T. Tani and S. Yamada, "Estimating user interruptibility by measuring table-top pressure," in *Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 1707–1712.

[93] J. Ho and S. S. Intille, "Using context-aware computing to reduce the perceived burden of interruptions from mobile devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2005, pp. 909–918.

[94] S. Mathan, S. Whitlow, M. Dorneich, P. Ververs, and G. Davis, "Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context," in *Proceedings of the 4th International Conference of the Augmented Cognition Society*, 2007.

[95] B. P. Bailey and S. T. Iqbal, "Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management," *ACM Transactions on Computer-Human Interaction*, vol. 14, no. 4, pp. 1–28, 2008.

[96] D. Chen, J. Hart, and R. Vertegaal, "Towards a physiological model of user interruptability," in *Human-Computer Interaction – INTERACT 2007*, vol. 4663, 2007, pp. 439–451.

[97] T. Fritz, E. M. Huang, G. C. Murphy, and T. Zimmermann, "Persuasive technology in the real world: A study of long-term use of activity sensing devices for fitness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2014, pp. 487–496. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557383

[98] R. Levandovski, E. Sasso, and M. P. Hidalgo, "Chronotype: a review of the advances, limits and applicability of the main instruments used in the literature to assess human phenotype," *Trends in psychiatry and psychotherapy*, vol. 35, no. 1, pp. 3–11, 2013.

[99] N. Hunn, "The market for smart wearables," WiFore Wireless Consulting, 2014.